Multi-Agent Preference-Based Reinforcement Learning

Data Mining & Quality Analytics Open Seminar

2026.05.22

김 정 인

발표자 소개

❖ 김정인 (Jung In Kim)

- Data Mining & Quality Analytics Lab
- Ph.D. Student (2021.09 ~)
- 지도 교수: 김성범 교수님

❖ 관심 연구 분야

- Deep Reinforcement Learning

❖ Contact

- Jungin_kim23@korea.ac.kr

목차

Introduction

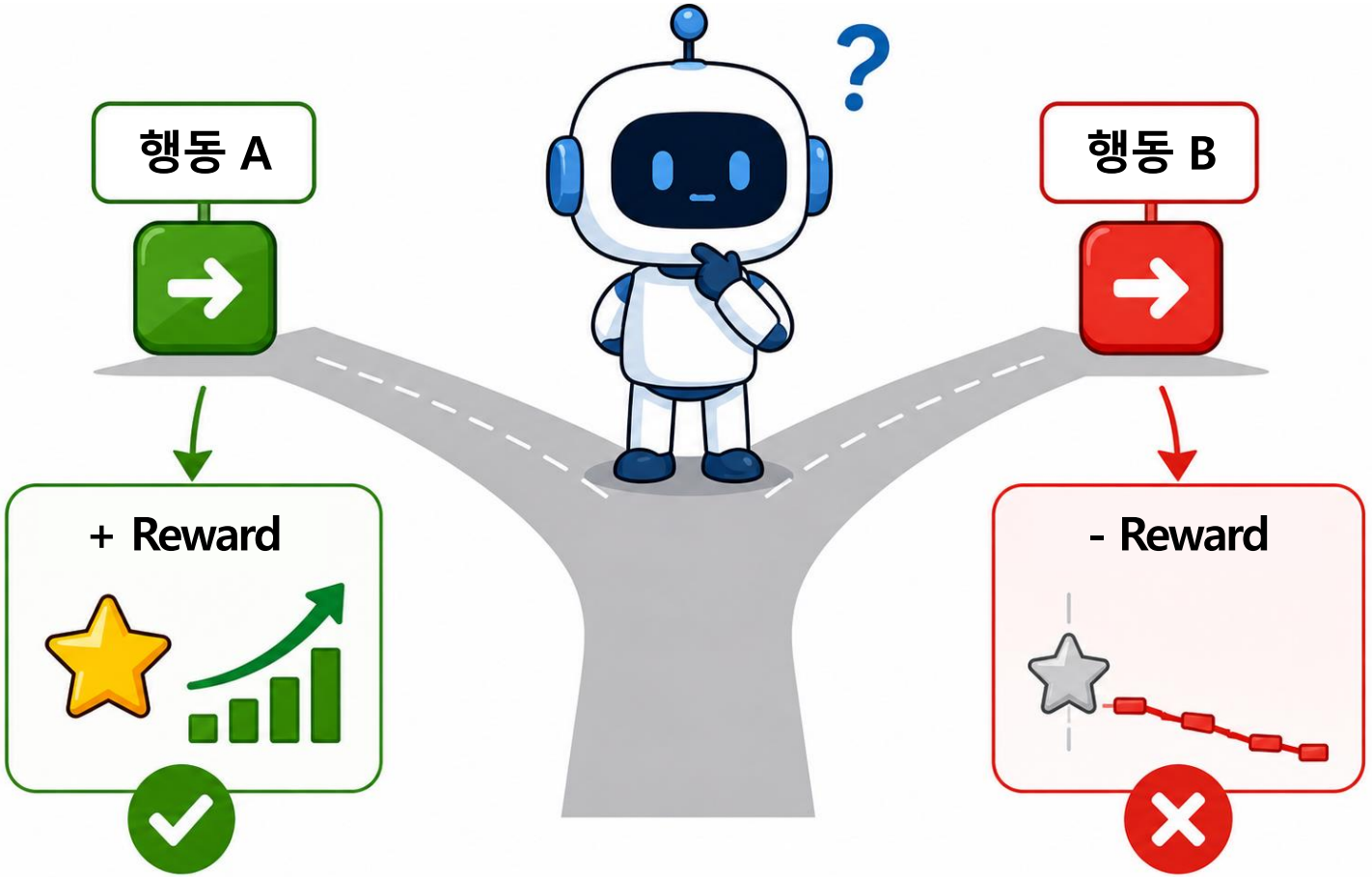
강화학습은 무엇일까?

❖ 순차적인 의사결정 문제에서 에이전트가 환경으로부터 받는 **누적 보상 값**을 최대화하는 정책을 학습하는 방법

Introduction

에이전트는 어떤 행동이 좋은 행동인지 어떻게 판단할까?

❖ 보상을 기준으로 행동의 좋고 나쁨을 판단함

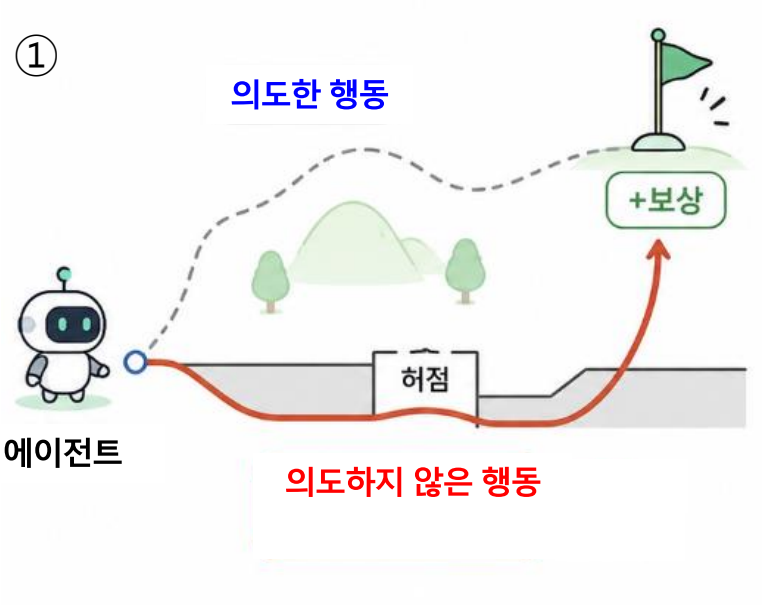


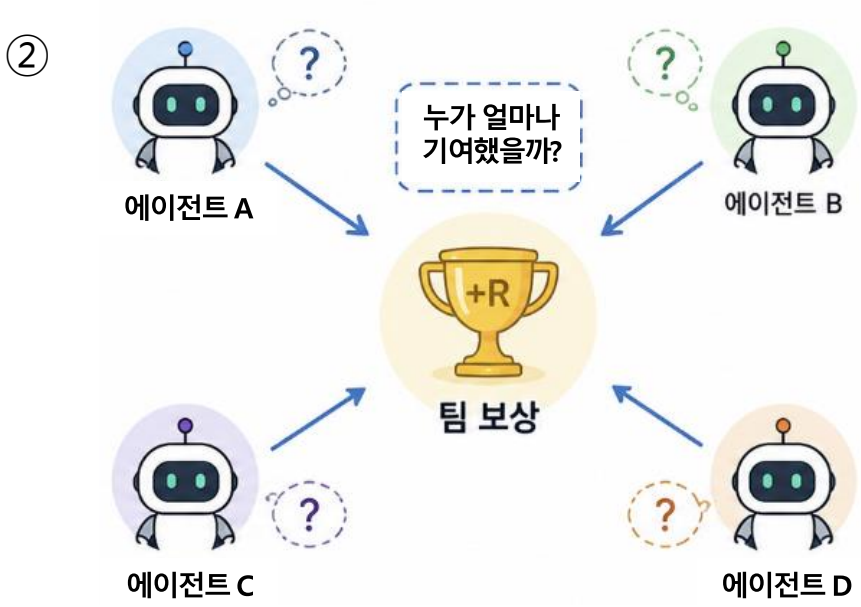
Introduction

보상 설계를 잘못하면 어떻게 될까?

① 잘못된 보상은 의도하지 않은 행동을 유도할 수 있음

② 특히, 다중 에이전트 강화학습에서는 여러 에이전트의 상호작용 때문에 보상 설계가 더 복잡함



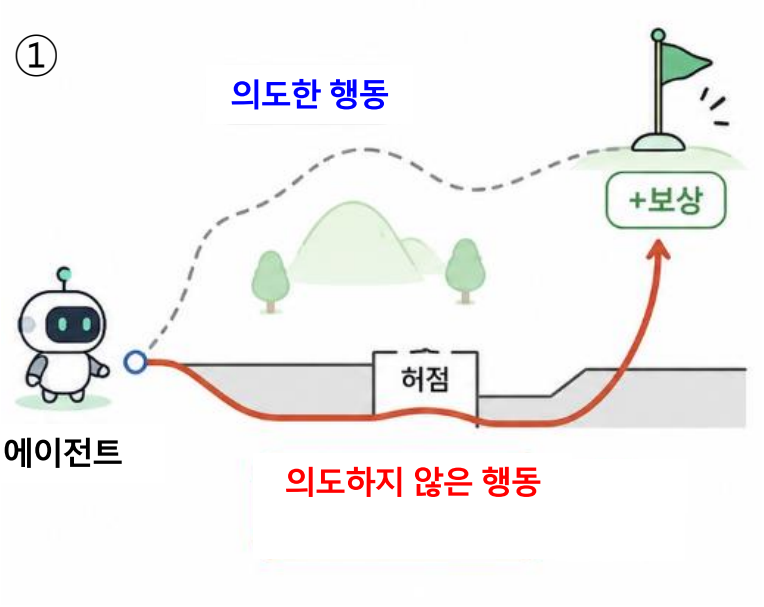


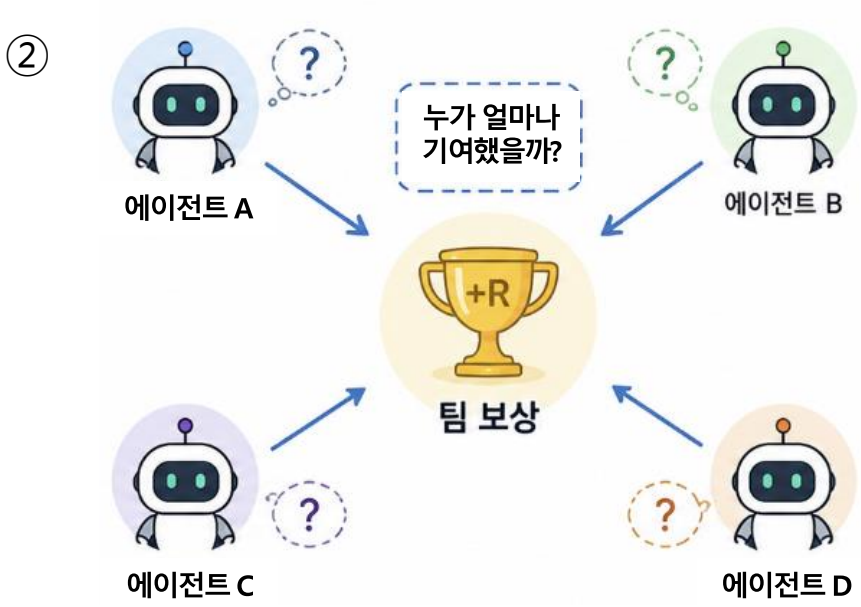
Introduction

보상 설계를 잘못하면 어떻게 될까?

① 잘못된 보상은 의도하지 않은 행동을 유도할 수 있음

② 특히, 다중 에이전트 강화학습에서는 여러 에이전트의 상호작용 때문에 보상 설계가 더 복잡함

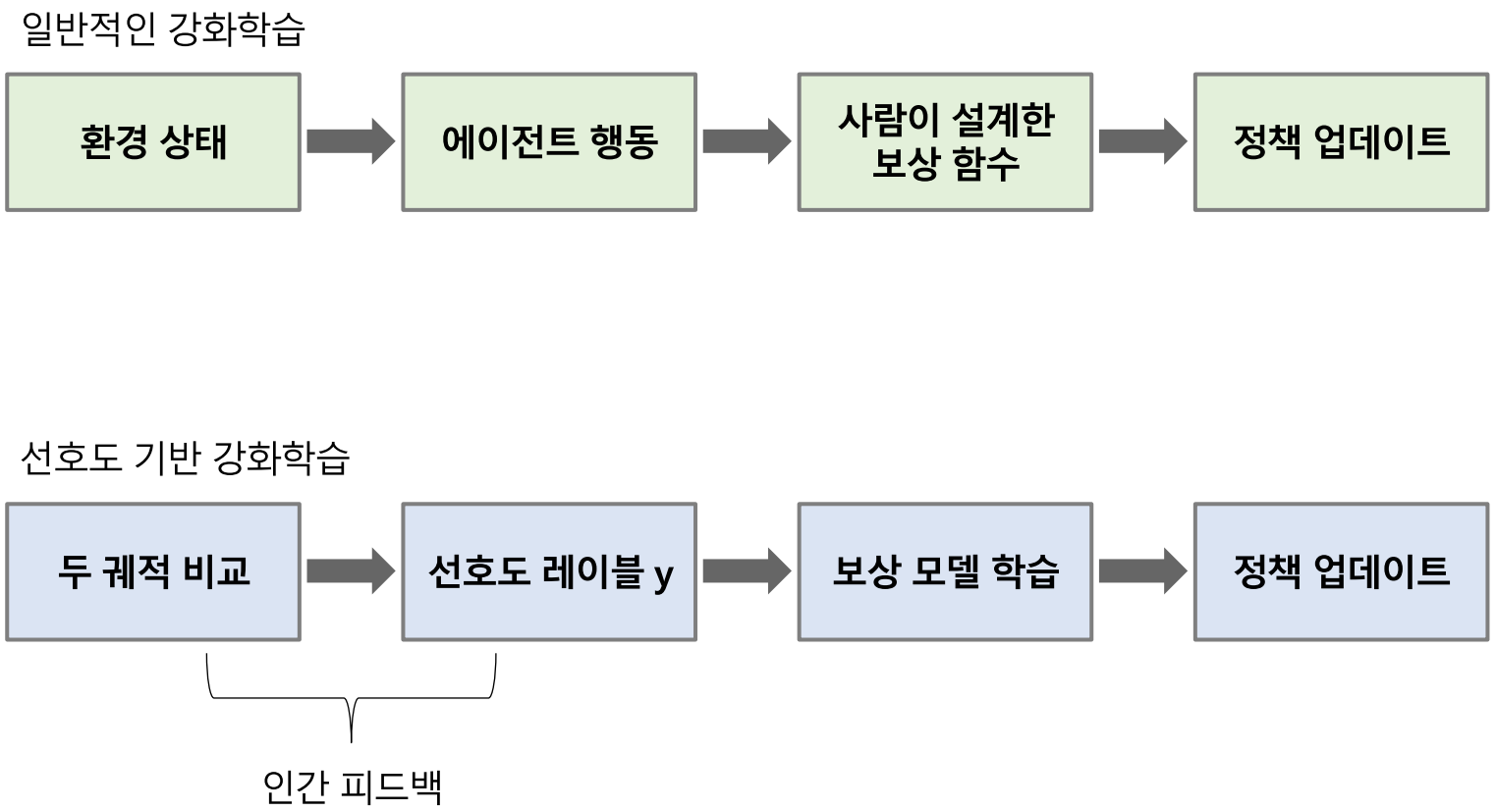




보상 함수를 사람이 직접 정확하게 정의하는 것은 굉장히 어려움

Introduction

선호도 기반 강화학습이란?



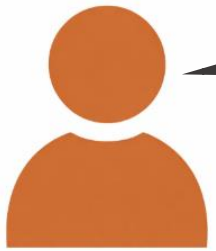
Introduction

인간의 선호도가 반영된 데이터 셋은 어떻게 만들지?

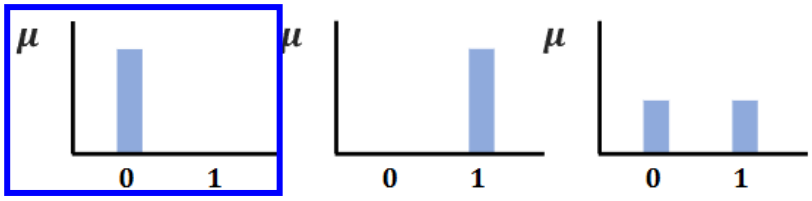
σ^0



σ^1



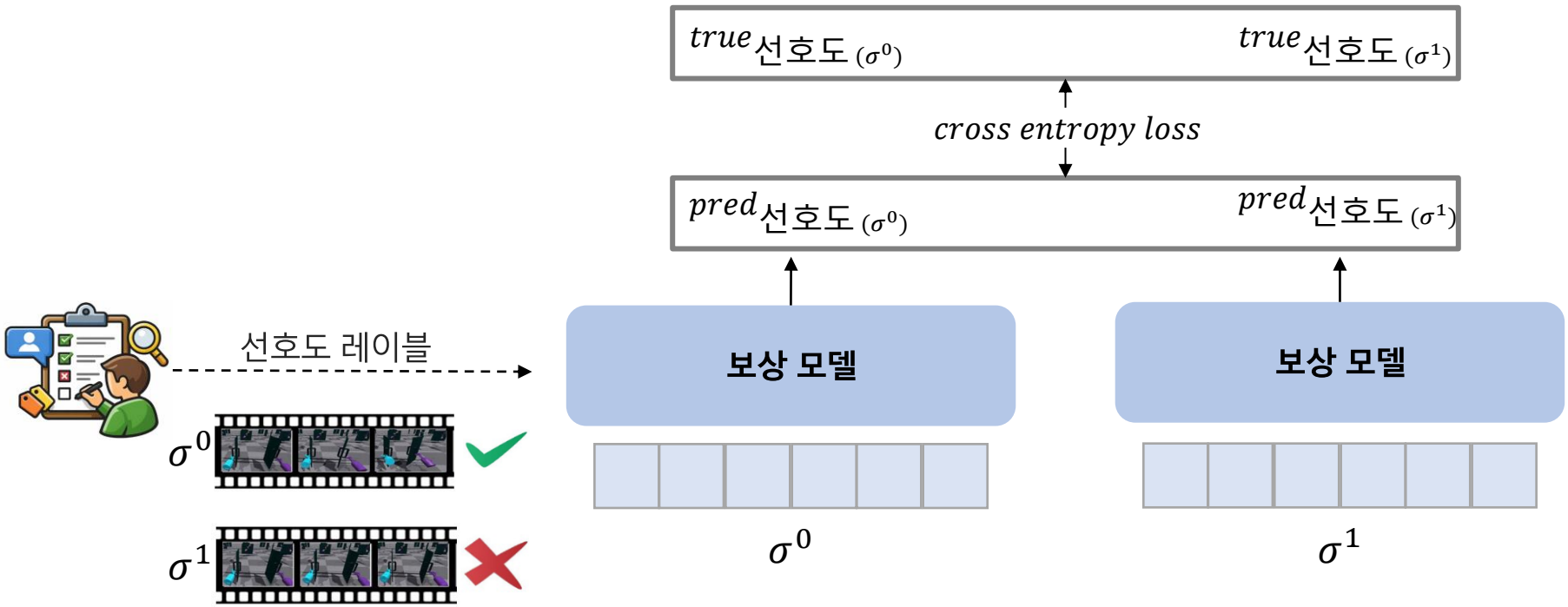
σ^0이 더 괜찮아 보여



Introduction

선호도 데이터 셋으로 보상 모델은 어떻게 학습하지?

- ❖ 각 데이터에 대한 선호도를 예측하고, 실제 정답 간의 cross entropy loss로 보상 모델을 학습



Methods

MARL + PbRL?

❖ 두 논문은 모두 PbRL을 MARL에 확장하지만, 문제를 바라보는 관점이 다름

MAPT

2024 AAAI

Reward model이
시간/협력 의존성을
잘 반영하는가?

Reward modeling 개선

AMADPO

2025 AAMAS

좋은 agent 행동을
선호 segment와 비선호
segment 모두에서 활용할
수 있는가?

Preference-behavior mismatch 해결

Methods

MAPT (2024, AAAI)

❖ 문제 상황

- 기존 단일 에이전트 PbRL을 MARL에 적용 시, **MARL의 시간적·협력적 의존성을 충분히 반영**하지 못함
  - ➢ 대부분 상태-행동 쌍에 동일한 보상 부여 → 탐색 다양성 제한 → 협력 학습 제한
- 핵심 아이디어: **cascaded transformer** 기반 보상 모델로 두 종류의 **global dependency**를 모델링

The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Decoding Global Preferences: Temporal and Cooperative Dependency Modeling in Multi-Agent Preference-Based Reinforcement Learning

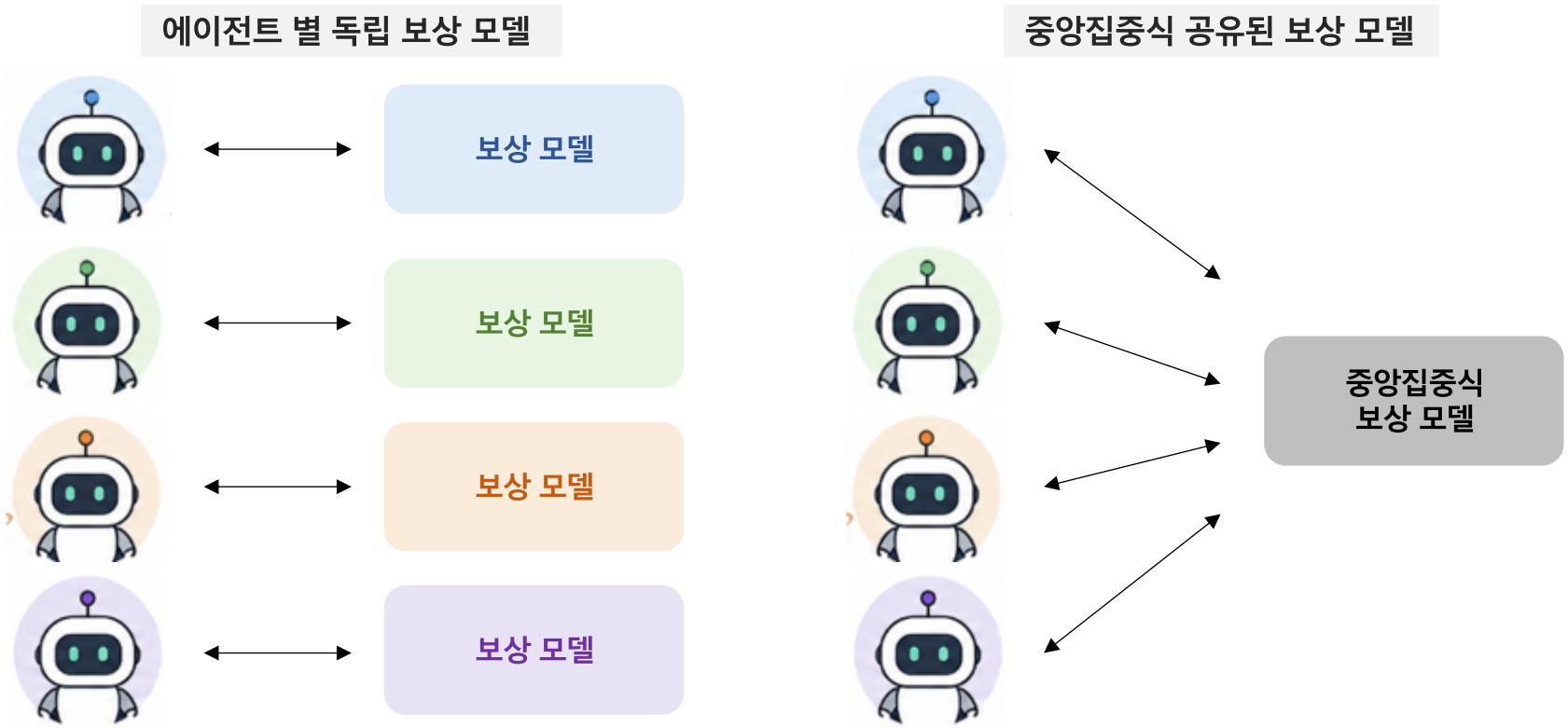
Tianchen Zhu[1], Yue Qiu[1], Haoyi Zhou[2, 3], Jianxin Li[1, 2]
[1]School of Computer Science and Engineering, Beihang University
[2]Zhongguancun Laboratory, Beijing, China
[3]School of Software, Beihang University
{zhutc,qiuyue,zhouhy,lijx}@act.buaa.edu.cn

Methods

MAPT (2024, AAAI)

❖ **기존 PbRL을 그대로 MARL에 어떻게 적용할까?**

- 에이전트 별 독립 보상 모델: 에이전트 협력을 놓칠 수 있음
- 중앙집중식 공유된 보상 모델: 각 에이전트의 기여도 분배가 어려움

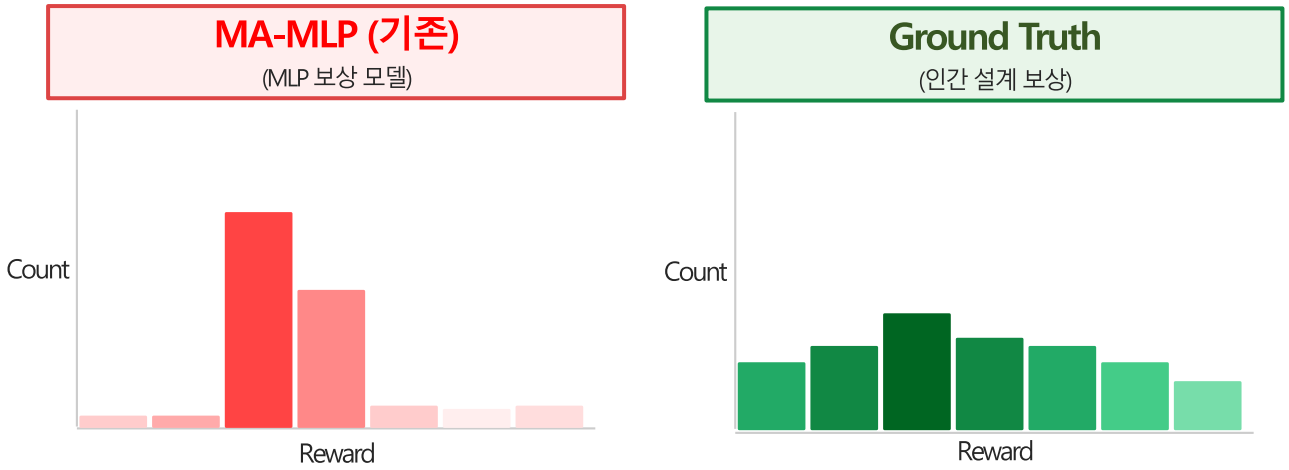


Methods

MAPT (2024, AAAI)

❖ **기존 PbRL을 그대로 MARL에 적용하면 발생하는 문제? Reward uniformity 문제**

- 여러 상태-행동 쌍에 비슷한 보상을 부여하는 문제가 발생함
- 결과적으로 다양한 행동 탐색과 협력 전략 학습이 어려워짐



선호도 → 전체 trajectory 평가
하지만, MARL에서는 "누가, 언제" 기여했는지가 중요

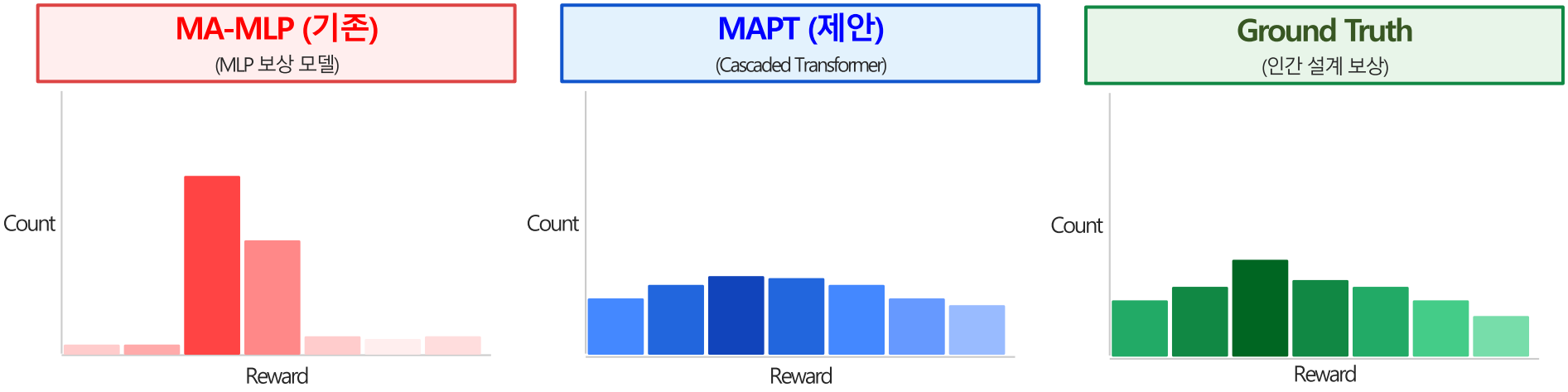
Methods

MAPT (2024, AAAI)

❖ **이러한 문제를 어떻게 개선해야할까?**

선호도 → 전체 trajectory 평가
하지만, MARL에서는 "누가, 언제" 기여했는지가 중요

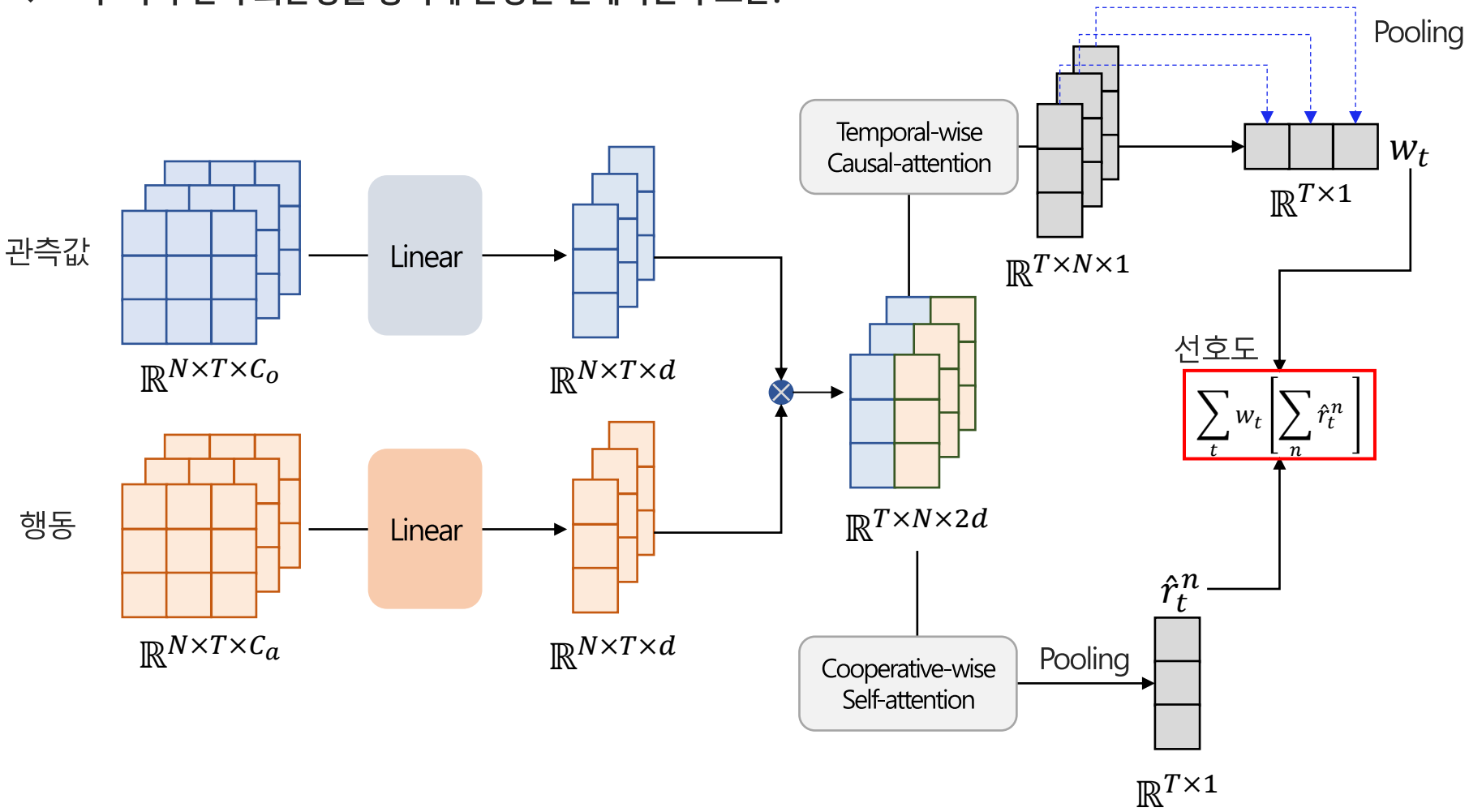
MAPT 해결 방향:
시간적(Temporal) + 협력적(Cooperative) 두 가지 전역 의존성 동시 포착 → **에이전트·타임스텝별 차별화된 보상 분포 학습**



Methods

MAPT (2024, AAAI)

❖ 두 가지 전역 의존성을 동시에 반영한 전체적인 구조는?



Methods

MAPT (2024, AAAI)

❖ **협력적 의존성 반영 과정은?**

$h_t^1, \cdots, h_t^N$
(N개 에이전트 임베딩 벡터)

↓ 선형 변환

$Q^n, K^n, V^n \in \mathbb{R}^d$

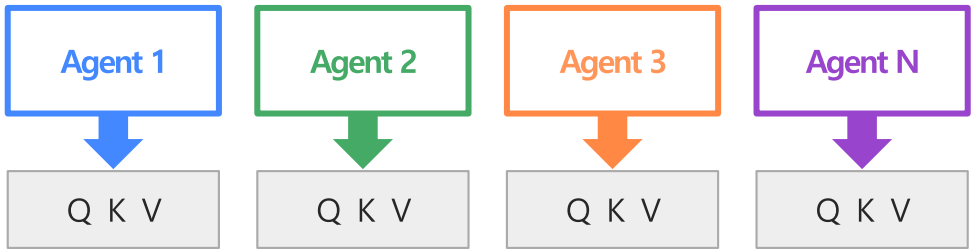
↓ Self-Attention

$$\hat{r}_t^n = softmax\left(\{\langle Q^n, K^k \rangle\}_{k=1}^{N}\right)_t^n \cdot V_t^n$$

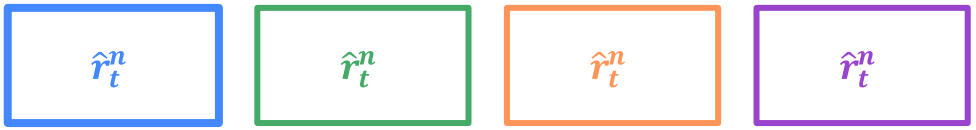
↓ 평균 풀링 (AGG)

$$\hat{r}_t = \left(\frac{1}{N}\right) \sum \hat{r}_t^n$$

N개 에이전트 Self-Attention 시각화



출력: 각 에이전트의 협력 반영 개별 보상



AGG (평균 풀링): $\hat{r}_t = \left(\frac{1}{N}\right) \sum \hat{r}_t^n$ → 집합 보상(Collective Reward)

Methods

MAPT (2024, AAAI)

❖ 시간적 의존성 반영 과정은?

입력

$\hat{r}_t,\ h_t = \frac{1}{N}\sum h_t^n$

↓ 선형 변환

$Q_t \leftarrow h_t, K_t \leftarrow \hat{r}_t^{1:N}, V_t \leftarrow \hat{r}_t$

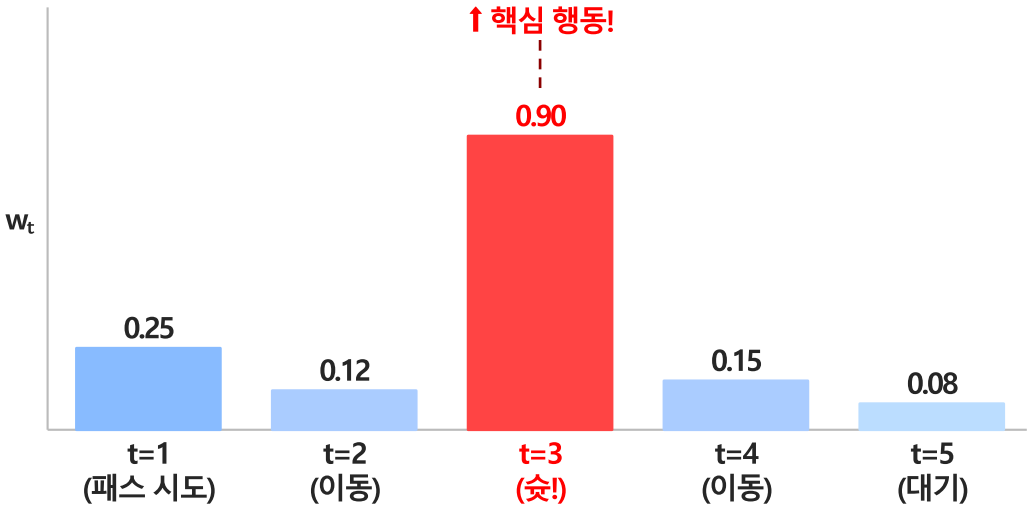
↓ Causal Self-Attention

$$x_k = \sum_{k=1}^{T} softmax\left(\{\langle Q_k, K_{t'}\rangle\}_{t'=1}^{T}\right)_t \cdot \hat{r}_t$$

↓ 가중 합산

$$\frac{1}{T}\sum_{t=1}^{T} x_k = \sum_{t=1}^{T} w_t \hat{r}_t$$

타임스텝별 중요도 가중치 w_t 시각화



최종 보상: $\sum_{t=1}^{T} w_t \hat{r}_t$

중요한 순간(슛, 패스 성공 등)에 높은 가중치 w_t

→ 선호도가 어느 시점 행동에서 비롯됐는지 파악 가능 (Temporal Credit Assignment)

Methods

MAPT (2024, AAAI)

❖ **시간적, 협력적 의존성을 고려한 MAPT가 효과가 있을까?**

- Scripted teacher 선호도 기반으로 선호도 기반 데이터 셋 확보
- **SMAC / GRF (50,000개), MaMuJoCo / Bi-Dexhands (30,000개)**

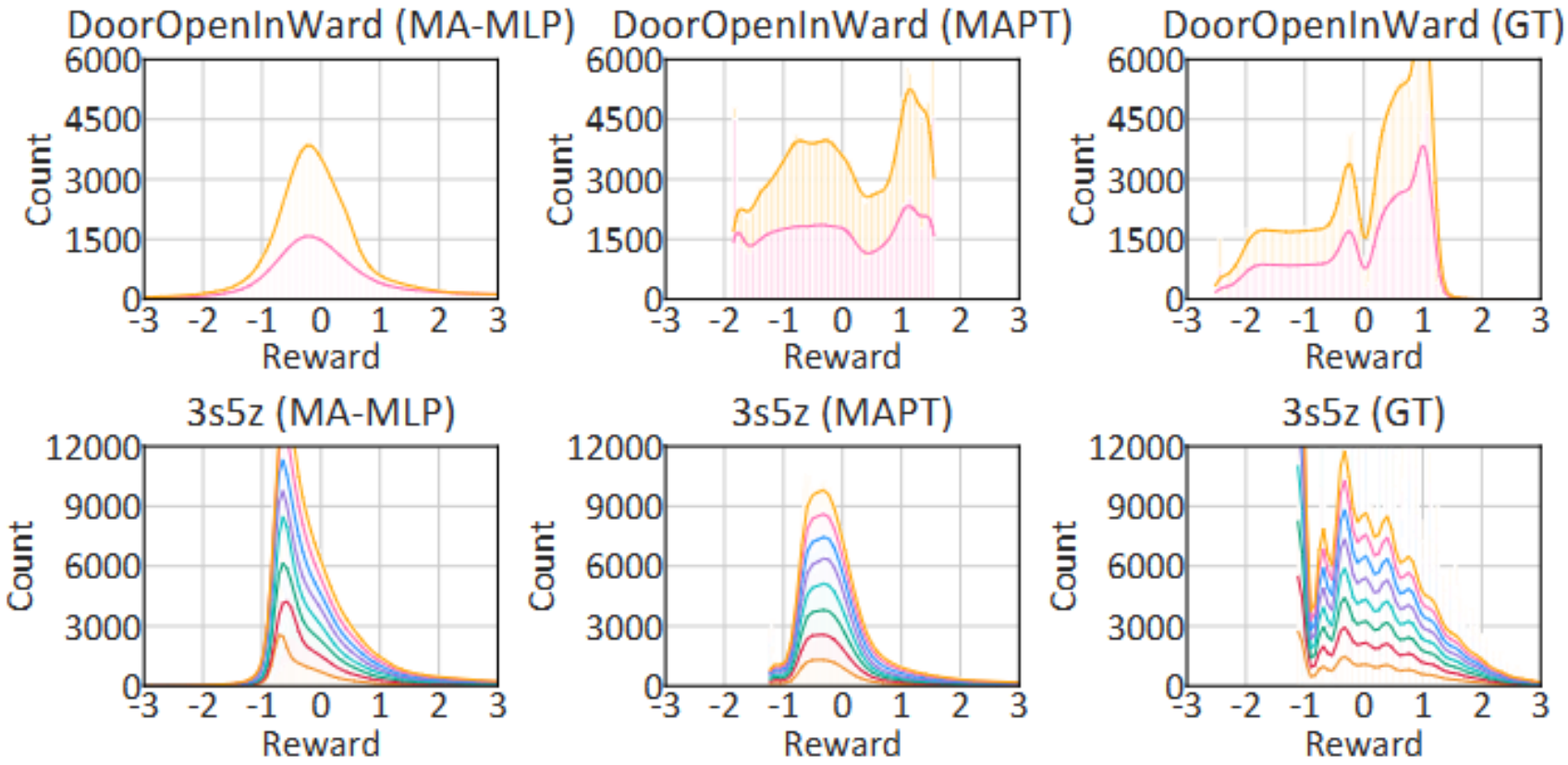
Tasks	Training with task rewards	Training with demonstations	Training with preferences			
	MAT	OMAR	MA-MLP	MA-LSTM	MA-Transformer	MAPT (Ours.)
3m	20.0±0.00	19.82±0.02	0.66±0.01	19.87±0.11	0.87±0.03	**19.98±0.05**
3s5z	20.00±0.02	19.87±0.11	1.77±0.17	14.52±0.49	5.86±0.28	**19.59±0.42**
6h vs 8z	19.78±0.07	18.33±0.15	9.57±0.14	9.76±0.15	6.36±0.17	**17.05±0.53**
MMM2	20.52±0.09	15.99±0.22	0.80±0.05	1.39±0.08	1.89±0.16	**3.68±0.10**
3 vs 1	4.89±0.02	4.55±0.34	0.72±0.06	1.31±0.03	1.06±0.02	**3.78±0.07**
counter attack	4.77±0.16	1.14±0.18	2.98±0.08	0.83±0.06	0.24±0.06	**4.21±0.10**
pass and shoot	4.83±0.11	2.72±0.58	2.20±0.06	1.90±0.03	0.66±0.06	**4.76±0.04**
CatchOver	25.32±0.88	16.85±1.21	8.62±0.34	**25.34±1.59**	4.75±0.13	25.12±0.64
DoorOpenInward	402.13±0.44	114.47±34.31	224.96±31.32	242.42±33.92	171.23±14.53	**372.60±11.38**
DoorOpenOutward	440.17±2.46	113.62±12.85	64.95±4.46	123.76±12.88	28.88±7.85	**228.08±10.98**
DoorCloseOutward	981.82±0.43	818.76±2.43	515.81±36.61	737.67±25.29	492.45±28.05	**786.70±26.03**
HalfCheetah 6×1	4483.95±74.75	4088.93±165.67	-88.75±11.62	1132.20±116.15	1317.90±147.77	**2423.50±128.33**

Methods

MAPT (2024, AAAI)

❖ **Reward uniformity 문제가 개선됐을까?**

- 각 선은 에이전트를 의미함



Methods

MAPT (2024, AAAI)

❖ 시간적, 협력적 의존성을 위한 셀프 어텐션이 보상 모델에 미치는 영향은?

Tasks	MAPT	MAPT-TPA	MAPT-CPA
OpenIn.	**292.16±11.38**	23.43±0.95	182.13±13.68
CloseOut.	**786.70±26.03**	718.16±22.38	782.37±24.05
3s5z	**19.59±0.42**	8.32±0.38	9.09±0.10
6h vs 8z	**17.05±0.53**	13.26±0.31	5.56±0.04

Methods

AMADPO (2025, AAMAS)

❖ 문제 상황

- Preference-Behavior mismatch를 주장 → 원인: **Global-local 불일치 + 두 단계 학습 불안정성**
- 핵심 아이디어: **보상 모델을 사용**하지 않고 **에이전트별 추가 로컬 선호도 레이블** 사용

Research Paper Track AAMAS 2025, May 19 – 23, 2025, Detroit, Michigan, USA

Offline Multi-Agent Preference-Based Reinforcement Learning with Agent-aware Direct Preference Optimization

Qian Kou[1]
Xi'an Jiaotong University[2]
Xi'an, China
xjtukouqian@stu.xjtu.edu.cn

Mingyang Li[1]
Xi'an Jiaotong University[2]
Xi'an, China
limingyang@stu.xjtu.edu.cn

Zeyang liu[1]
Xi'an Jiaotong University[2]
Xi'an, China
zeyang.liu@xjtu.edu.cn

Long Qian
Xi'an Jiaotong University[2]
Xi'an, China
qianlongym@stu.xjtu.edu.cn

Zhuoran Chen
Xi'an Jiaotong University[2]
Xi'an, China
zhuoran.chen@xjtu.edu.cn

Lipeng Wan
Xi'an Jiaotong University[2]
Xi'an, China
wanlipeng77@xjtu.edu.cn

Xingyu Chen[3]
Xi'an Jiaotong University[2]
Xi'an, China
chenxingyu_1990@xjtu.edu.cn

Xuguang Lan[3]
Xi'an Jiaotong University[2]
Xi'an, China
xglan@mail.xjtu.edu.cn

Methods

AMADPO (2025, AAMAS)

❖ **Preference-Behavior mismatch란?**

- 선호도 데이터는 "어떤 trajectory segment (σ^0 or σ^1)가 더 좋은가?"를 전역(global) 수준에서만 알려줌
- 하지만, 정책 학습 결과가 실제 선호도와 일치하지 않을 수 있음

즉, 선호도 라벨이 정책이 학습해야 할 개별 행동 신호와 일치하지 않는 문제를 의미

Methods

AMADPO (2025, AAMAS)

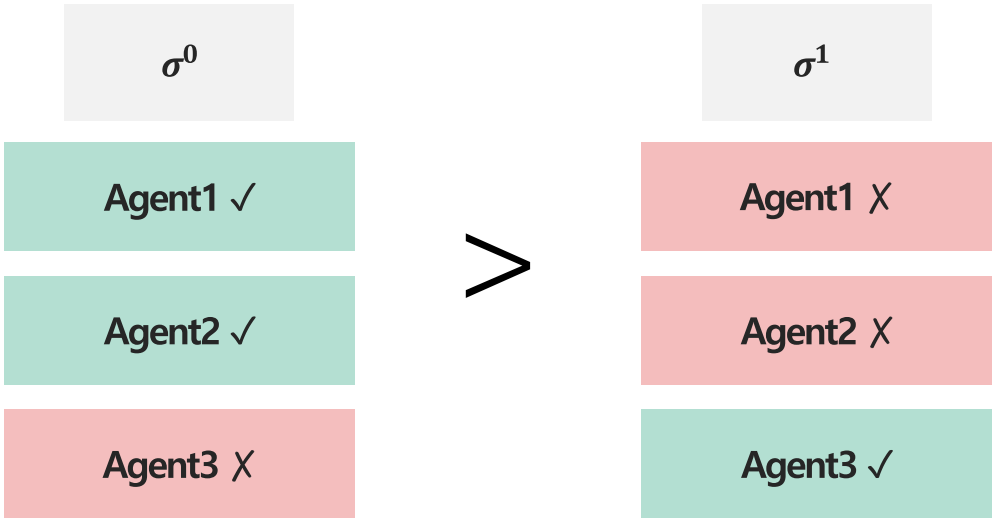
❖ **Preference-Behavior mismatch란?**

- 선호도 데이터는 "어떤 trajectory segment (σ^0 or σ^1)가 더 좋은가?"를 전역(global) 수준에서만 알려줌
- 하지만, 정책 학습 결과가 실제 선호도와 일치하지 않을 수 있음

즉, 선호도 라벨이 정책이 학습해야 할 개별 행동 신호와 일치하지 않는 문제를 의미

❖ **Preference-Behavior mismatch가 왜 발생하는가?**

- 원인 1: global-local preference inconsistency



Agent3는 덜 선호된 σ^1에서 더 좋아!

Methods

AMADPO (2025, AAMAS)

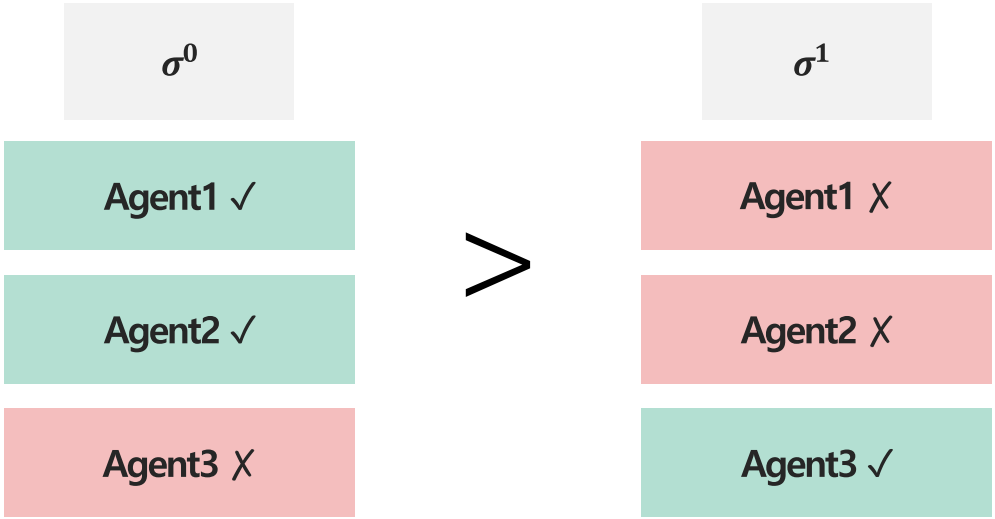
❖ **Preference-Behavior mismatch란?**

- 선호도 데이터는 "어떤 trajectory segment (σ^0 or σ^1)가 더 좋은가?"를 전역(global) 수준에서만 알려줌
- 하지만, 정책 학습 결과가 실제 선호도와 일치하지 않을 수 있음

즉, 선호도 라벨이 정책이 학습해야 할 개별 행동 신호와 일치하지 않는 문제를 의미

❖ **Preference-Behavior mismatch가 왜 발생하는가?**

- 원인 1: global-local preference inconsistency



따라서, agent-level 선호도가 필요함

Methods

AMADPO (2025, AAMAS)

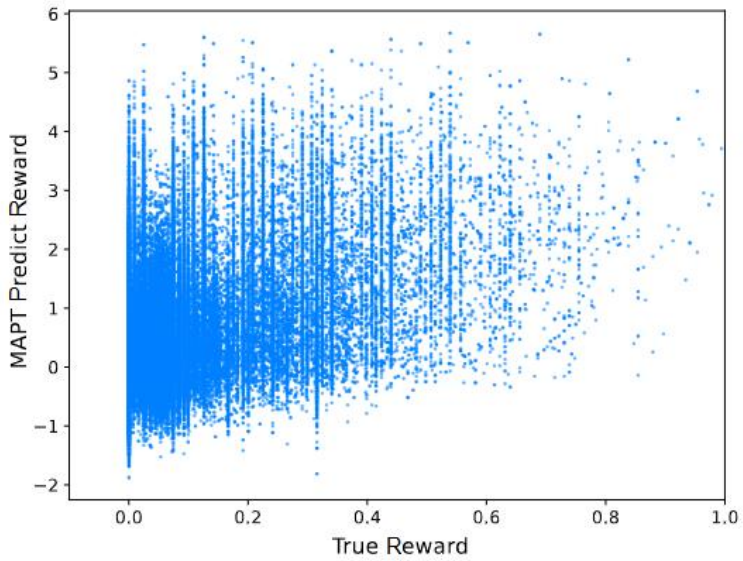
❖ Preference-Behavior mismatch란?

- 선호도 데이터는 "어떤 trajectory segment (σ^0 or σ^1)가 더 좋은가?"를 전역(global) 수준에서만 알려줌
- 하지만, 정책 학습 결과가 실제 선호도와 일치하지 않을 수 있음

즉, 선호도 라벨이 정책이 학습해야 할 개별 행동 신호와 일치하지 않는 문제를 의미

❖ Preference-Behavior mismatch가 왜 발생하는가?

- 원인 2: 보상 모델과 정책을 따로 학습하는 two-stage 구조의 불안정성



MAPT가 예측한 보상과 실제 보상 사이의 불일치

Methods

AMADPO (2025, AAMAS)

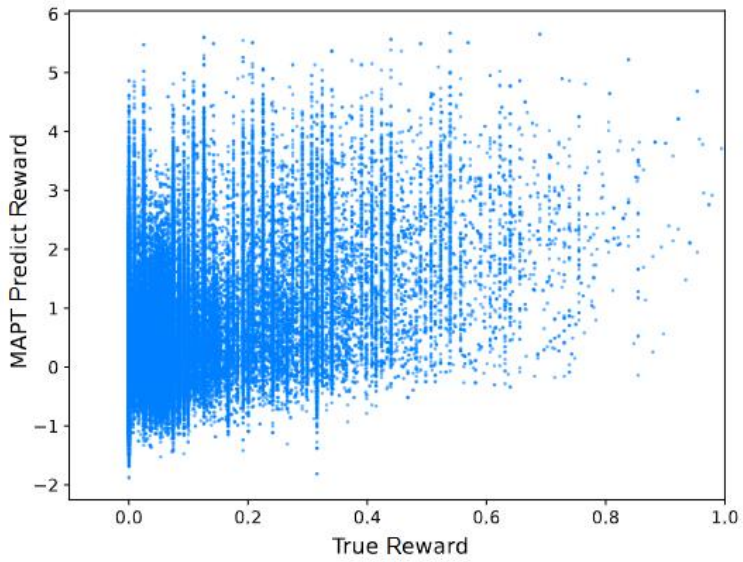
❖ Preference-Behavior mismatch란?

- 선호도 데이터는 "어떤 trajectory segment (σ^0 or σ^1)가 더 좋은가?"를 전역(global) 수준에서만 알려줌
- 하지만, 정책 학습 결과가 실제 선호도와 일치하지 않을 수 있음

즉, 선호도 라벨이 정책이 학습해야 할 개별 행동 신호와 일치하지 않는 문제를 의미

❖ Preference-Behavior mismatch가 왜 발생하는가?

- 원인 2: 보상 모델과 정책을 따로 학습하는 two-stage 구조의 불안정성



따라서, 보상 모델을 중간 목표로 두지 않고 preference objective로 정책을 직접 업데이트하고자 함

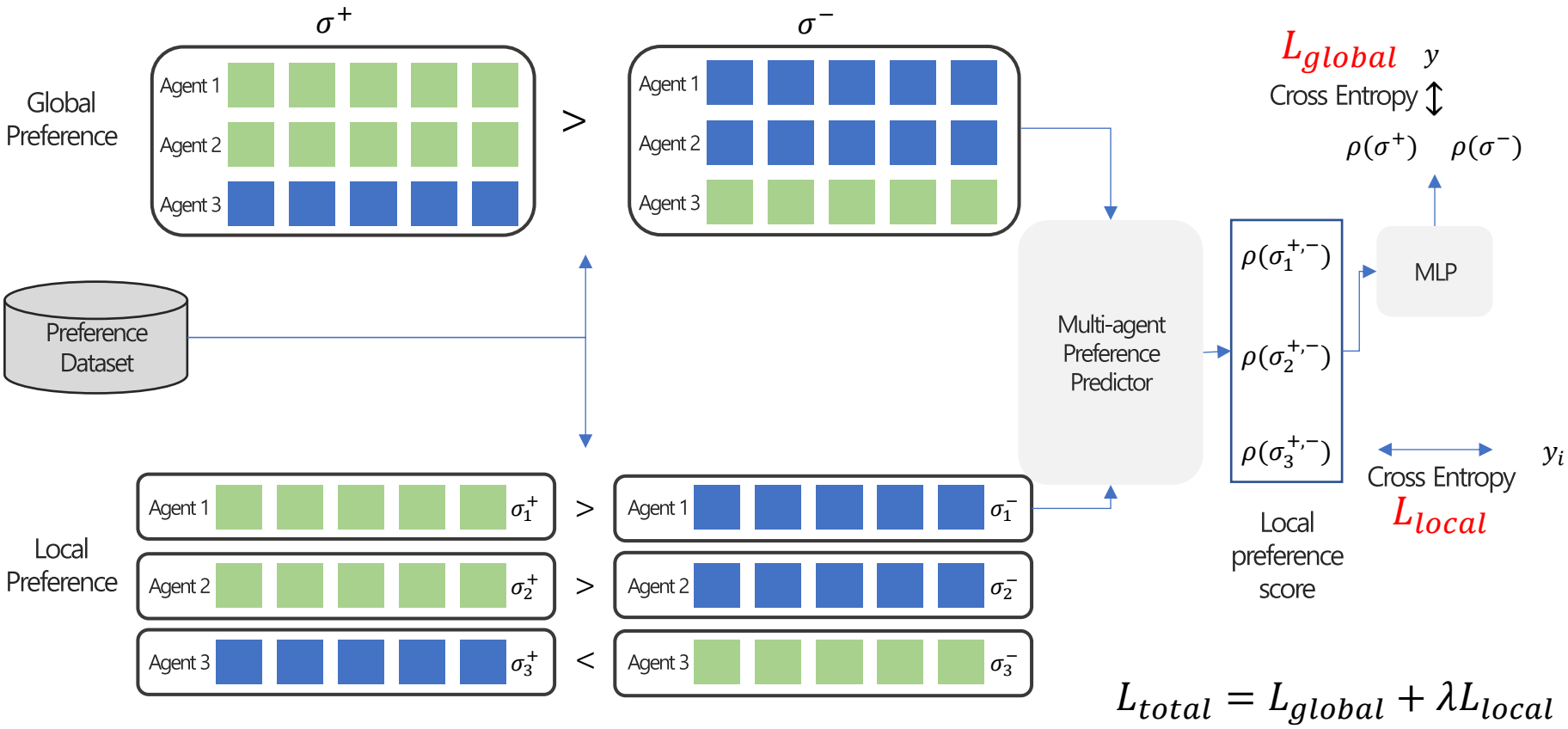
Methods

AMADPO (2025, AAMAS)

❖ **Global-local 불일치와 보상 모델 사용의 불안정성 문제를 해결한 전체적인 구조는?**

- Stage 1: Global과 local 선호도를 모두 정확하게 예측하는 Multi-agent preference predictor 학습

Stage 1: Multi-agent Preference Predictor Training



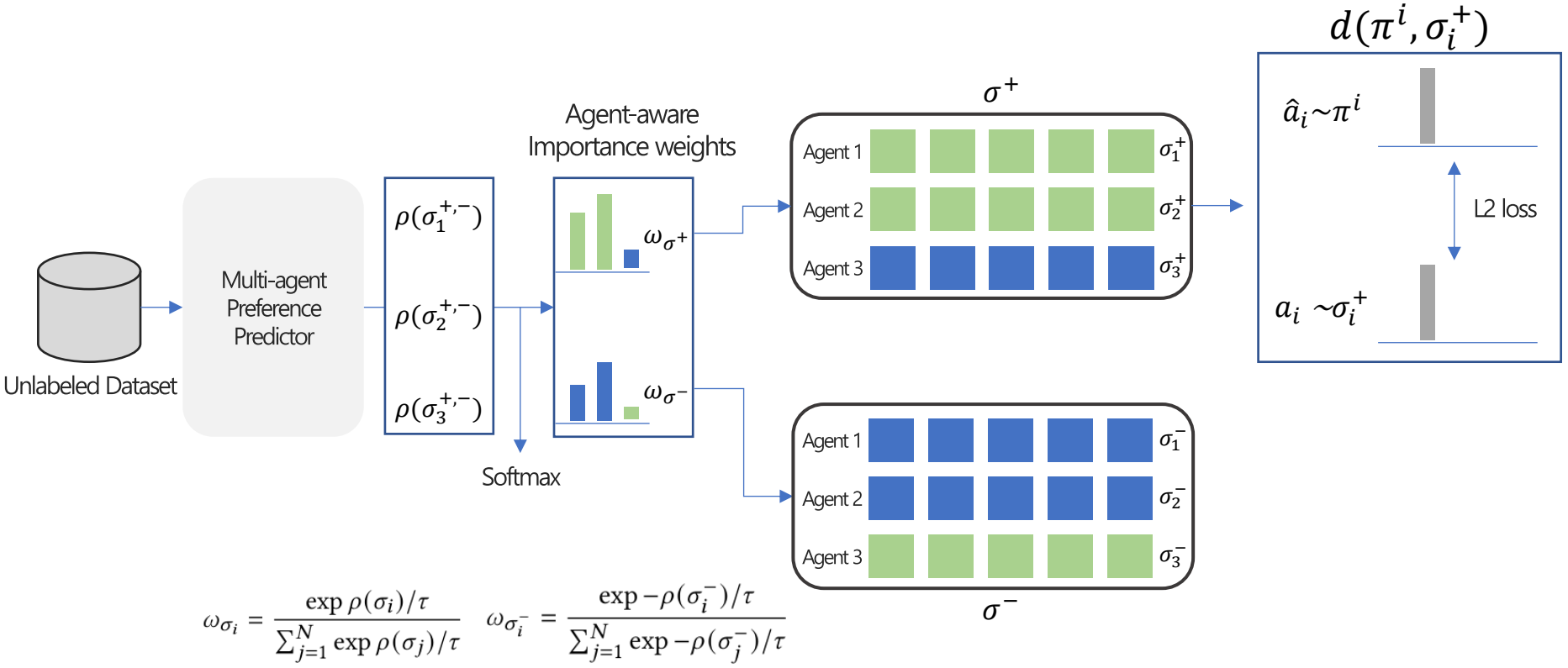
Methods

AMADPO (2025, AAMAS)

❖ **Global-local 불일치와 보상 모델 사용의 불안정성 문제를 해결한 전체적인 구조는?**

- Stage 2: 학습된 multi-agent preference predictor를 사용해서 실제 정책을 최적화하는 단계

Stage 2: Agent-aware Direct Preference Optimization



$$\omega_{\sigma_i} = \frac{\exp \rho(\sigma_i)/\tau}{\sum_{j=1}^{N} \exp \rho(\sigma_j)/\tau} \qquad \omega_{\sigma_i^-} = \frac{\exp -\rho(\sigma_i^-)/\tau}{\sum_{j=1}^{N} \exp -\rho(\sigma_j^-)/\tau}$$

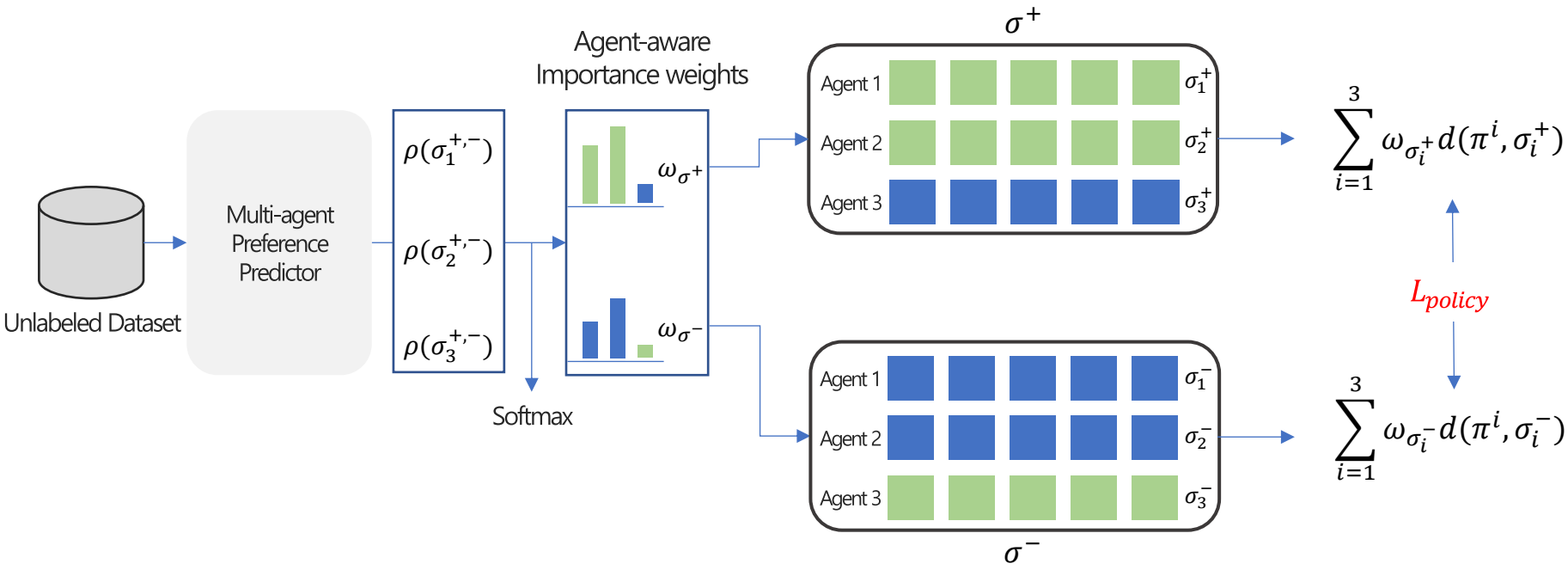
Methods

AMADPO (2025, AAMAS)

❖ **Global-local 불일치와 보상 모델 사용의 불안정성 문제를 해결한 전체적인 구조는?**

- Stage 2: 학습된 multi-agent preference predictor를 사용해서 실제 정책을 최적화하는 단계

Stage 2: Agent-aware Direct Preference Optimization



$$L_{policy} = -E_{(\sigma^+,\sigma^-)\sim D_{pref}}[log \frac{exp - \sum_{i=1}^{3} \omega_{\sigma_i^+} d(\pi^i, \sigma_i^+)}{exp - \sum_{i=1}^{3} \omega_{\sigma_i^+} d(\pi^i, \sigma_i^+) + \lambda exp - \sum_{i=1}^{3} \omega_{\sigma_i^-} d(\pi^i, \sigma_i^-)}]$$

Methods

AMADPO (2025, AAMAS)

❖ **Global-local 불일치성과 보상 모델 사용의 불안정성을 고려한 AMADPO가 과연 효과가 있을까?**

- 비교 방법론

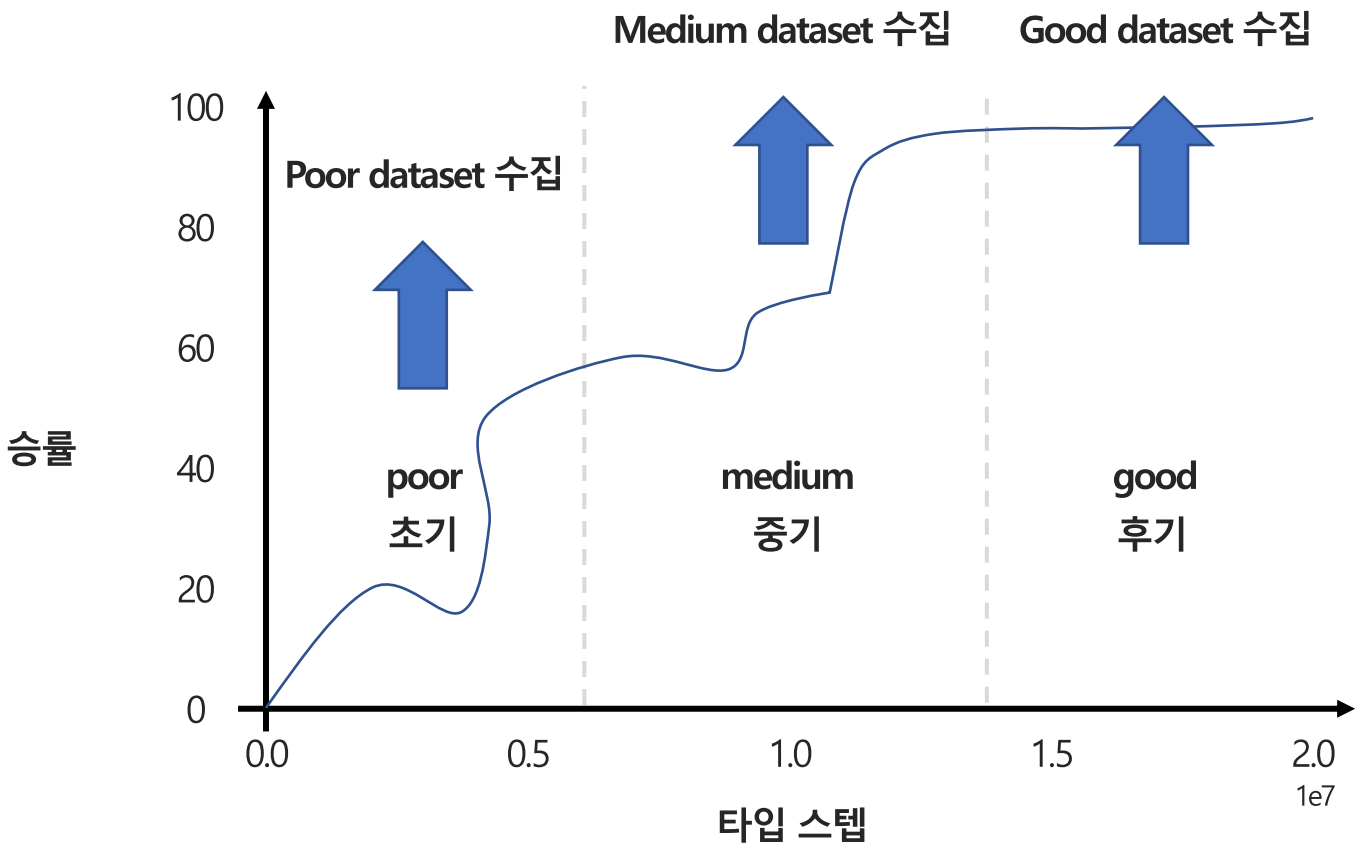
Offline MARL	Two-stage PbRL	DPO 계열
BC (Behavior Cloning)	ICQ + MLP	CPL
ICQ with GT reward	ICQ + LSTM	DPPO
	ICQ + PT / MAPT	

Methods

AMADPO (2025, AAMAS)

❖ **Global-local 불일치성과 보상 모델 사용의 불안정성을 고려한 AMADPO가 과연 효과가 있을까?**

- 데이터 셋 Low, Medium?

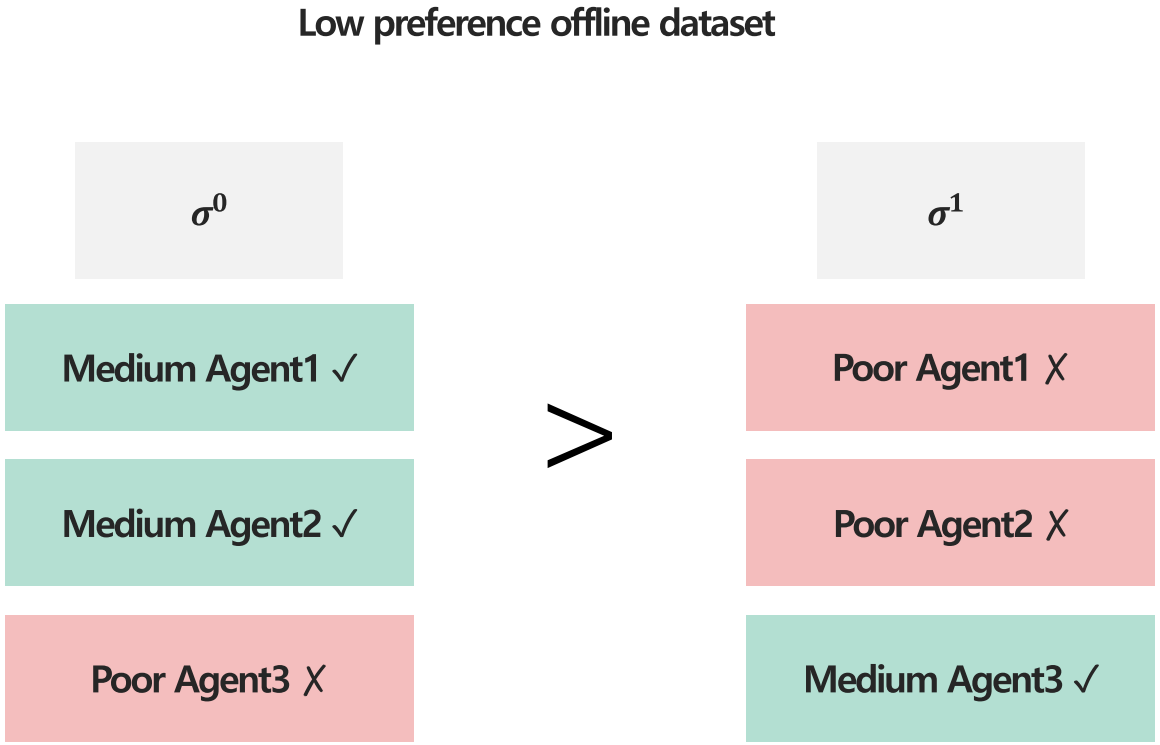


Methods

AMADPO (2025, AAMAS)

❖ **Global-local 불일치성과 보상 모델 사용의 불안정성을 고려한 AMADPO가 과연 효과가 있을까?**

- 데이터 셋 Low, Medium?



Methods

AMADPO (2025, AAMAS)

❖ **Global-local 불일치성과 보상 모델 사용의 불안정성을 고려한 AMADPO가 과연 효과가 있을까?**

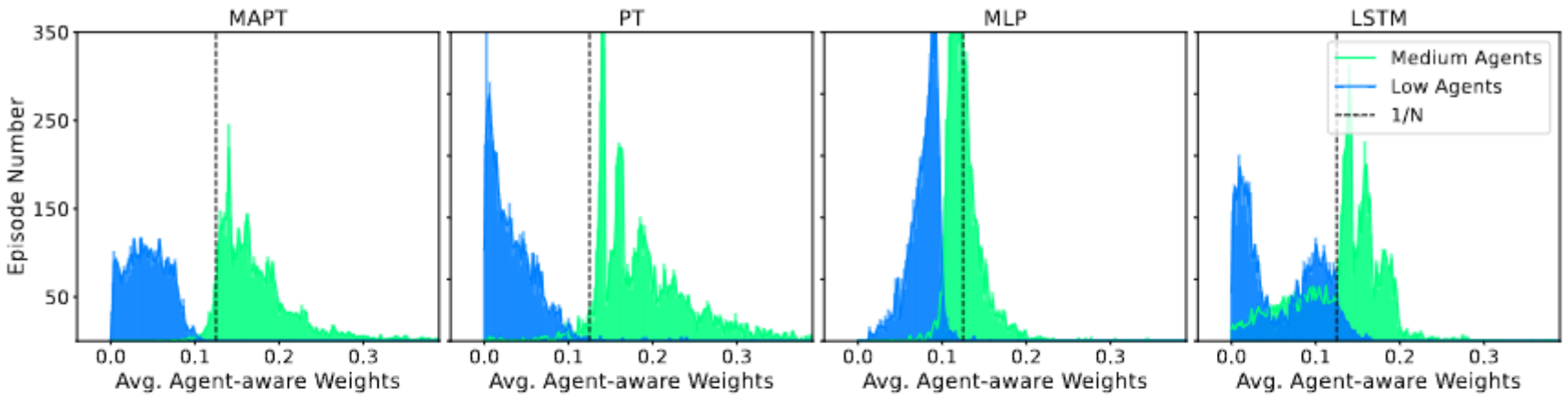
- 보상 모델 기반 방법보다 direct preference optimization의 안정성이 강조됨

Difficulty	Task	Dataset	BC	ICQ	ICQ+MLP	ICQ+LSTM	ICQ+PT	ICQ+MAPT	CPL	DPPO	**Ours**
Easy	3m	Low	3.74±1.31	5.79±1.19	0.14±0.06	4.16±0.72	5.02±0.25	4.29±1.03	5.12±1.38	5.72±3.26	**8.79±1.47**
		Medium	4.79±0.57	7.68±1.26	0.16±0.09	10.6±1.49	9.72±1.68	**11.1±2.6**	6.67±0.9	6.11±3.87	8.56±0.93
	8m	Low	3.31±0.15	5.00±1.66	0.17±0.03	0.17±0.06	0.17±0.09	0.17±0.07	3.64±0.26	2.94±0.24	**9.83±0.51**
		Medium	4.43±0.33	11.54±1.46	2.7±0.23	2.7±0.67	2.7±0.54	2.7±0.05	7.41±1.45	10.28±2.82	**14.97±1.46**
	2s3z	Low	4.56±0.78	8.13±0.61	5.26±0.57	8.44±0.95	6.89±0.75	0.52±0.38	6.52±2.12	3.11±1.74	**9.07±0.33**
		Medium	5.77±0.95	**13.9±0.78**	0.51±0.39	11.13±0.29	12.45±0.46	0.5±0.37	9.5±1.14	3.57±1.11	**13.04±0.32**
	1c3s5z	Low	6.78±1.29	7.88±0.16	2.32±0.21	2.82±0.15	3.3±0.8	3.68±0.35	**9.44±0.84**	7.74±0.26	**8.47±0.93**
		Medium	8.76±0.41	**17.7±1.11**	4.69±0.18	6.79±4.31	3.07±0.34	8.08±5.25	12.7±0.35	11.04±0.13	**12.62±0.26**
Hard	3s_vs_5z	Low	5.4±0.33	6.66±0.65	2.74±0.48	2.26±0.63	4.7±0.21	2.69±0.63	**6.67±1.31**	5.52±0.24	**5.95±1.23**
		Medium	8.79±1.61	11.5±1.34	1.68±0.65	1.82±0.66	1.03±0.22	1.65±1.45	8.96±4.62	9.47±2.67	**12.53±1.60**
	5m_vs_6m	Low	5.38±1.29	5.28±0.41	0.32±0.21	4.02±0.15	5.3±0.8	4.68±0.35	3.71±0.23	2.68±0.32	**6.22±0.24**
		Medium	4.83±0.31	5.36±0.63	0.71±0.24	2.87±0.24	3.07±0.34	5.78±0.35	4.5±0.34	3.69±0.16	**6.42±0.42**
	8m_vs_9m	Low	5.38±1.29	6.39±1.18	0.32±0.21	8.02±1.39	5.3±0.8	6.9±1.08	7.02±1.38	7.46±1.33	**8.55±1.14**
		Medium	4.83±0.31	9.57±0.58	0.35±0.24	3.21±0.25	8.75±0.26	9.54±0.48	7.52±0.37	3.04±0.48	**11.88±0.58**
	10m_vs_11m	Low	4.77±0.47	8.92±0.16	0.14±0.06	9.47±0.99	7.62±0.29	**9.71±0.18**	7.82±0.19	6.54±0.34	**8.94±0.26**
		Medium	5.06±0.37	8.95±0.36	0.13±0.09	9.63±0.13	9.6±0.65	**10.1±0.32**	8.13±0.23	10.06±0.32	**9.13±0.16**
Super Hard	MMM2	Low	2.66±0.27	6.37±0.36	0.25±0.20	4.16±0.15	6.48±0.45	6.10±0.59	4.55±0.24	3.67±0.22	**6.58±0.47**
		Medium	3.21±0.25	9.14±0.32	0.21±0.19	3.68±0.58	6.74±0.86	5.80±0.43	7.42±1.45	8.36±1.66	**9.54±0.88**
	6h_vs_8z	Low	5.79±0.13	**7.22±0.64**	5.31±0.59	5.59±0.58	7.02±0.79	5.65±0.63	6.72±0.63	6.88±0.45	**6.97±0.25**
		Medium	3.21±0.25	**11.3±0.75**	1.21±0.56	2.63±0.52	5.56±0.45	1.30±0.69	8.69±0.67	6.62±0.26	**9.34±0.70**
	corridor	Low	5.33±0.22	6.69±0.50	1.44±0.52	1.66±0.62	1.62±0.84	6.05±0.42	6.42±0.53	4.52±0.18	**7.06±0.49**
		Medium	6.66±0.57	10.4±0.45	1.33±0.46	1.11±0.42	2.19±0.55	0.63±0.43	7.75±0.22	5.42±0.56	**13.7±0.33**
	3s5z_vs_3s6z	Low	5.62±0.33	7.74±0.19	1.49±0.26	0.56±0.22	8.02±0.23	2.69±2.53	6.56±0.19	4.92±0.43	**8.09±0.99**
		Medium	7.24±0.14	13.89±0.47	0.74±0.26	10.1±0.36	6.12±0.35	0.61±0.29	7.8±0.15	6.29±0.46	**14.14±0.47**

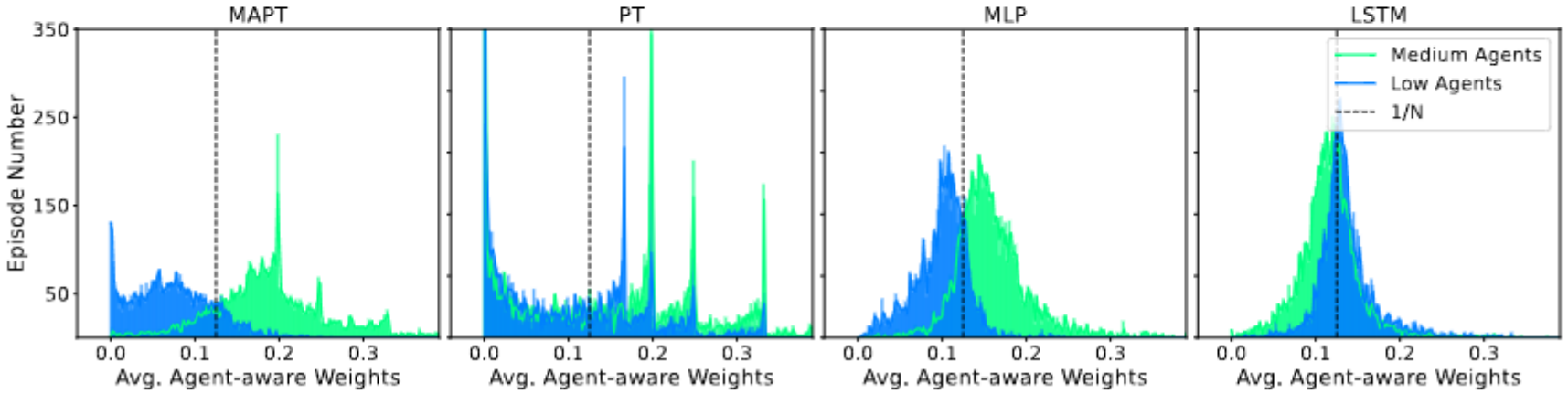
Methods

AMADPO (2025, AAMAS)

❖ Multi-agent preference predictor의 에이전트별 가중치는 어떻게 될까?



(a) Average agent-aware weights of Preference Predictors in Low quality MMM2 task.



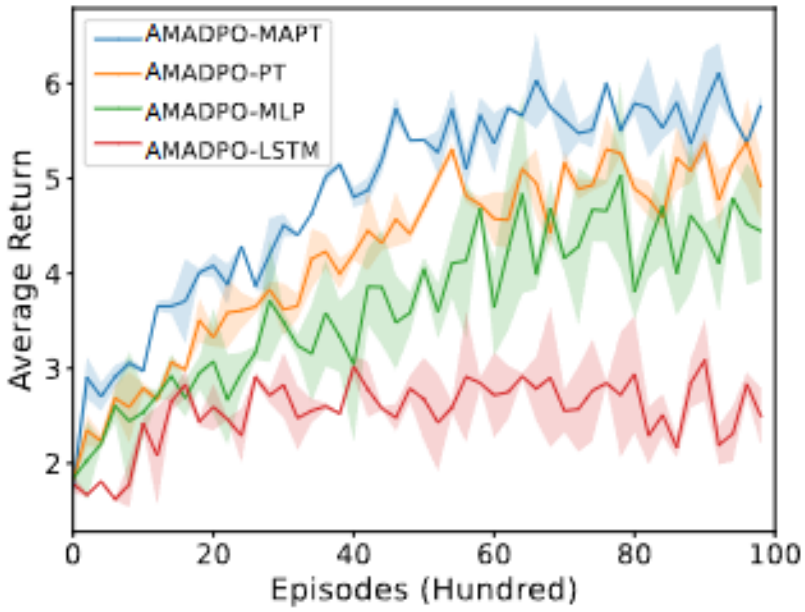
(b) Average agent-aware weights of Preference Predictors in Low quality 3s5z_vs_3s6z task.

Methods

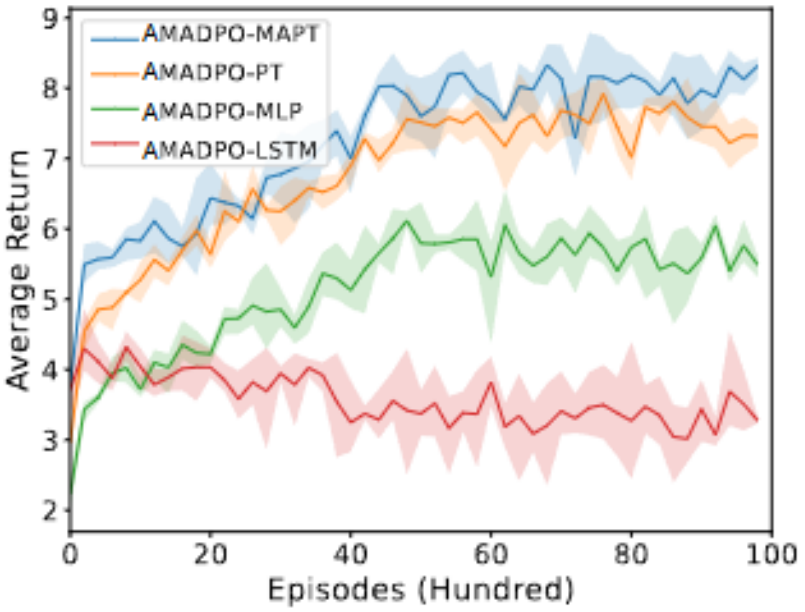
AMADPO (2025, AAMAS)

❖ **Multi-agent preference predictor의 종류에 따른 AMADPO의 성능은 어떻게 될까?**

- 에이전트별 가중치를 얼마나 잘 계산하는지에 따라 AMADPO의 최종 정책 학습 성능이 달라질 수 있음



(c) Learning curves of AMADPO in Low quality MMM2 task.



(d) Learning curves of AMADPO in Low quality 3s5z_vs_3s6z task.

Conclusion

- ❖ MAPT (2024, AAAI)
  - 시간적, 협력적 의존성을 반영한 보상 모델링의 필요성을 제시한 연구
- ❖ AMADPO (2025, AAMAS)
  - 전역 선호도만 사용 및 보상 모델의 불안정성으로 인해 발생하는 preference-behavior mismatch 문제를 해결한 연구

Reference

1. Zhu, T., Qiu, Y., Zhou, H., & Li, J. (2024, March). Decoding global preferences: Temporal and cooperative dependency modeling in multi-agent preference-based reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 15, pp. 17202-17210).
2. Kou, Q., Li, M., Liu, Z., Qian, L., Chen, Z., Wan, L., ... & Lan, X. (2025, May). Offline Multi-Agent Preference-Based Reinforcement Learning with Agent-aware Direct Preference Optimization. In Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (pp. 1181-1190).

Thank you